

7. A Personal View of Statistical Packages for Linear Regression

Jerome Vanclay

This module presents some personal observations and views on statistical packages for forestry research, with particular attention on suitability for regression analysis, and important activity in development of systems models. There are many packages able to assist statistical analyses of data, designed for different purposes and audiences. This module illustrates some of the features that are particularly important when choosing a package for linear regression analyses. Among these is the graphical abilities of a package, especially the ease with which data (and residuals) can be plotted, because a model may be judged by the residuals rather than by the R-squared statistic. However, preferences for particular features and approaches are personal, so provided that candidates have the functionality one needs, a package may be chosen on the basis of personal preference. If you don't like the one you have, or find that it cannot handle your data or your analysis, look for another package. There are many well designed and tested statistical packages available, so it should always be possible to find one that suits your needs and budget.

1. INTRODUCTION

Two-variable and multivariate analysis are important steps in fitting relationships for use in systems models. Many statistical packages for use on personal computers are available with regression capabilities, and there is great variation in range of capabilities, ease of use and cost. This is a personal overview of statistical packages used by the author, emphasizing the utility of the package for fitting curves to data using linear regression. It is not a comprehensive review, and does not consider expensive packages such as SPSS, SAS, and S-plus. Instead, it looks mainly at the free or cheap packages that do not require an annual license fee. Some basic concepts in regression analysis are first introduced, and then a number of packages with regression capabilities are reviewed – specifically Excel, CurveExpert, GLIM, ARC and ViSta.

2. SOME BASIC CONCEPTS IN STATISTICS

Let's begin by re-examining the principles underlying curve fitting with regression analysis. Figure 1 is a graph with six data points. I'm going to ask you to draw the

single straight line that best describes the trend evident in these points. Your line should be the free-hand equivalent of the least-squares approach used in regression analysis. Understanding the position of this line is fundamental to an understanding of regression analysis. Go ahead, and draw your line.

Examining the data graphically, and developing a mental image of the best fit are important first steps in model-building. As Forrest Young (2001) puts it in poetry,

*first, you see your data
for what they seem to be
then, you ask them for the truth
- are you what you seem to me?*

*you see with broad expanse
yet ask with narrowed power
you see and ask and see
and ask and see ... and ask ...*

*with brush you paint the possibilities
with pen you scribe the probabilities
for in pictures we find insight
while in numbers we find strength*

Forrest W. Young, 2001

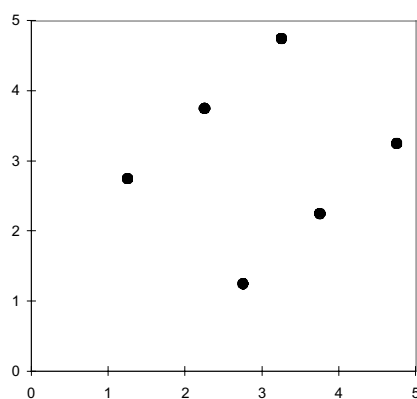


Figure 1. A scatter plot.

Now compare your line with those illustrated in Figure 2. Many people draw the first principal component (dotted line) rather than the least squares fit (solid line). They are minimizing the overall distance between the points and their line, but that's not what linear regression does. Linear regression assumes that the X-values (the numbers on the horizontal axis) are known with certainty, and that any errors are associated only with the Y-values. Thus linear regression minimizes the vertical distances between the points and the lines (Figure 3). In fact, it minimizes the square of those distances – that's why it's called the *least squares* fit, and why one of the measures of goodness-of-fit is the residual mean square (the average of the vertical lines squared). How did your line compare with those in Figure 2?

The vertical lines in Figure 3 are known as residuals. Examining the residuals is a useful way to judge the quality of a fit. And because the regression technique is based on the square of these residuals, you should take special note of large residuals, which may have a strong influence on the shape and position of the fitted line. Consult any standard text on regression analysis for a fuller explanation. The text by Cook and Weisberg (1994) is a helpful place to begin.

Many people judge the quality of a fit using a statistic known as R-squared. An R-squared of zero indicates that the fitted line is no better than a simple average, and an R-squared of one reveals a perfect fit. When R-squared is close to one, it may indicate a

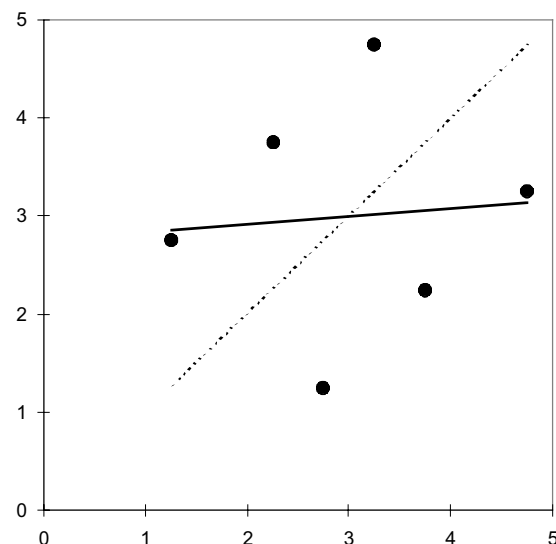


Figure 2. First principal component (dotted line) and least-squares fit (solid line) to the six data points (redrawn from Vanclay 1994, Figure 6.3)

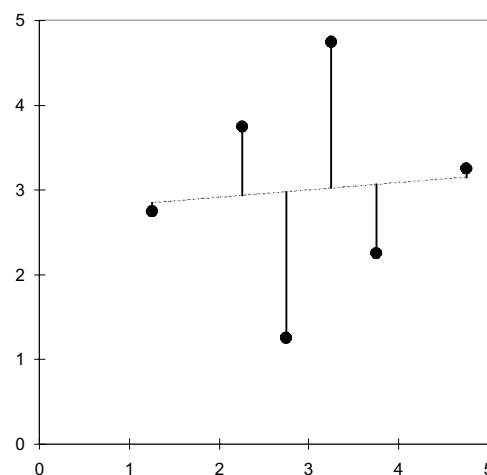


Figure 3. Regression analysis chooses the dotted line in such a way as to minimize the sum of the squares of the solid vertical lines

good fit, but it does not indicate if the fit is good enough.

Anscombe (1973) created four data sets that reveal some limitations of R-squared and emphasise the need to examine data graphically (Figure 4). The four data sets that he formulated have exactly the same linear regression estimates ($Y = 3.0 + 0.5 X$), exactly the same residual mean square (13.75) and exactly the same R-squared (0.667). However, despite these similar values, the quality of the fit varies greatly

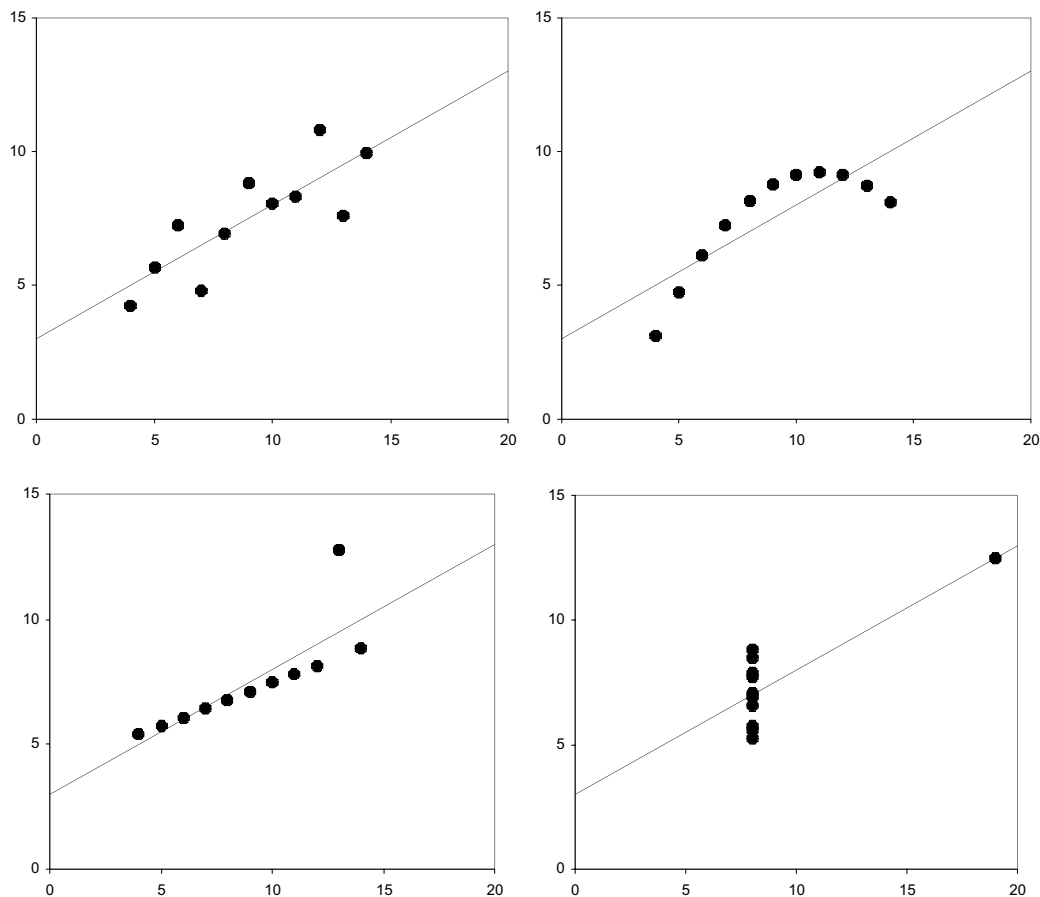


Figure 4. Four data sets constructed by Anscombe (1973) in which the R-squared is exactly the same (0.667), but in which the quality of the fit varies greatly

between the four data sets (Figure 4). They reveal 'pure error' (top left), an outlier (bottom left), use of the wrong model (a quadratic term may be needed; top right), and a case where the estimated trend relies entirely on a single point with high leverage (bottom right). Without further information, it is impossible to judge which of these models is suitable, but all except the first case warrant careful investigation.

A fuller discussion of this illustration was given by Anscombe (1973) and Weisberg (1985), but the importance of plotting both the data and the model remains obvious.

Because it is so important to plot the data and examine the points visually, I am going to consider only those statistical packages that have strong graphics capabilities. As Forrest Young (2001) says in his poem, it is *'... in pictures we find insight, while in numbers we find strength'*.

Five packages will be illustrated, and used to fit a simple linear regression ($Y = a + bX$) to a small set of data. The data are 10 pairs of numbers, with $X = 1, 2, 3, \dots, 10$ and $Y = X^2$. Fitting a straight line isn't particularly appropriate, but the R-squared is close to one (0.95). We contrast the insights we gain into this analysis as we examine the five packages – Excel, CurveExpert, GLIM, ARC and ViSta.

3. MICROSOFT EXCEL

Microsoft Excel is part of the Microsoft Office suite of programs, and is available on many computers. It's not my favourite package, but it is handy and sometimes may be the only statistical software available.

To perform regression analysis with Excel, click on 'Tools' and on 'Data Analysis'. Use the slider to find 'Regression', select and

click 'OK', and the computer screen should look like Figure 5. If the Tools menu doesn't include 'Data Analysis', it may be necessary to select the add-in: Click on 'Tools' and on 'Add-Ins', and make sure that the 'Analysis ToolPak' is selected. If you cannot find the 'Add-Ins', ask your computer support people to install the full version of Excel.

When performing a regression analysis with Excel, be sure to plot the raw data with the fitted model, and examine the residuals (Choose both the 'Residual Plots' and 'Line Fit Plots' in the regression dialogue box, Figure 5), and do not rely only on the R-squared and the F-test to judge the adequacy of a model.

Statistical procedures in Excel are not always reliable. McCullough and Wilson (1999) reported that the Statistical Reference Dataset of the American National Institute of Standards and Technology reveals problems with Excel's univariate summary statistics, analysis of variance, linear regression, non-linear regression and random number generation. The problems appear to be present in Excel 4, Excel 95, Excel 97 and Excel 2000. Thus

it may be advisable not to use Excel for complex or critical analyses.

4. CURVE EXPERT

CurveExpert has limited capabilities, fitting curves to (X, Y) pairs of data, but it is an easy way to fit an equation to data. CurveExpert automatically fits and compares 35 built-in regression models and up to 15 additional user-defined models comprising up to 19 parameters. These include both linear and non-linear regression as well as splines. There is no limit on the number of data points. Good documentation is available on the Help menu.

Data entry and analysis involves simply typing pairs of data (X, Y) into the built-in spreadsheet, and pressing control-f, and CurveExpert will establish the line of best fit, graph it, and display the corresponding equation (Figure 6). Graphs can be customized and copied onto the clipboard for use in other Windows applications.

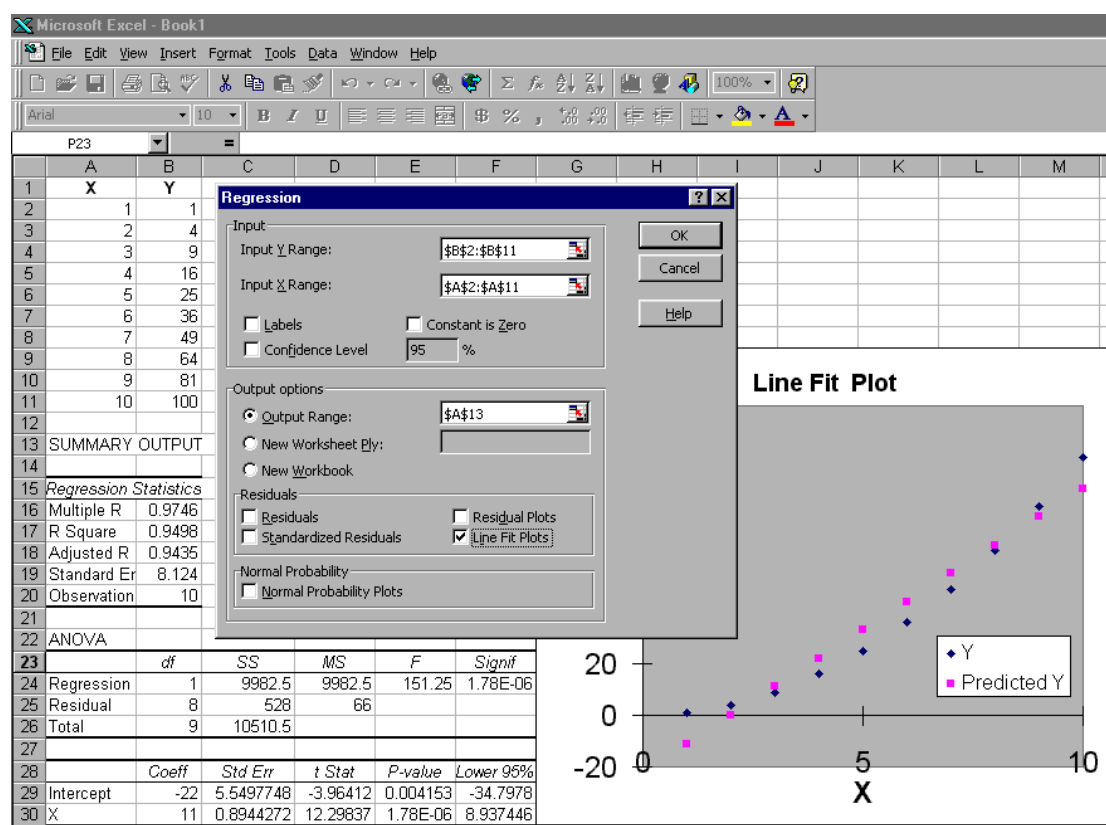


Figure 5. Image of a computer screen during a regression analysis with Excel

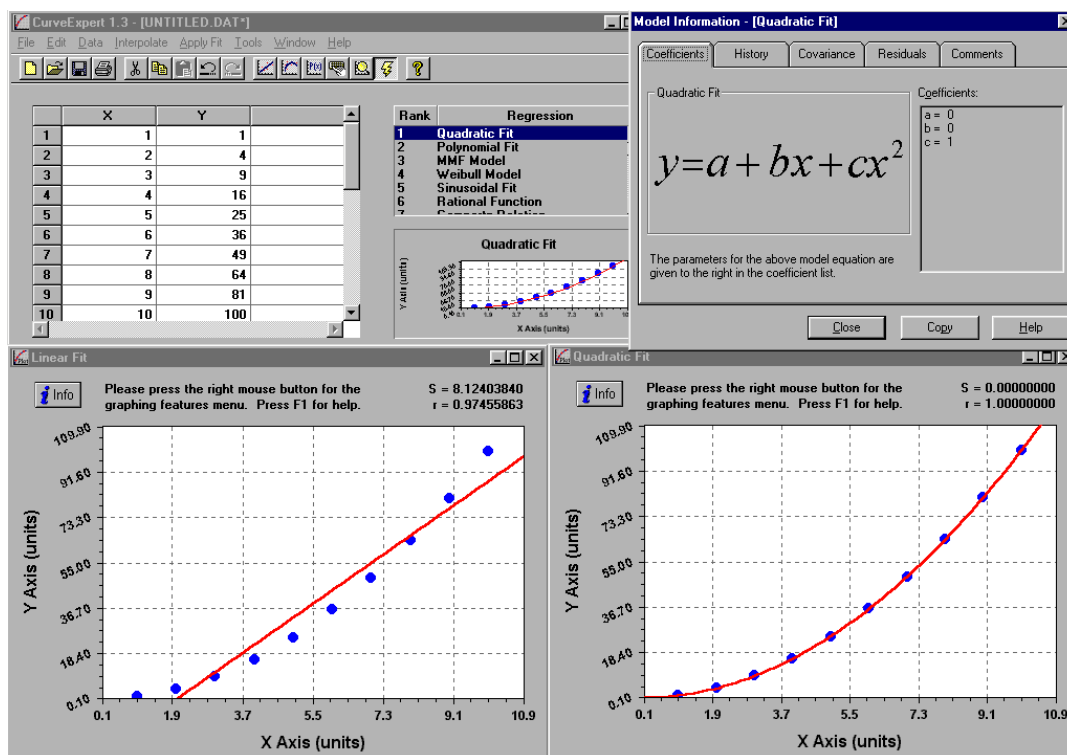


Figure 6. Image of computer screen while fitting a curve with CurveExpert

The major limitation of CurveExpert is that it can use only pairs of data. This is adequate if you want to build a simple yield table based only on stand age, but means that you cannot easily include site index or other explanatory variables.

CurveExpert is shareware developed by Daniel Hyams. It is available freely on the internet at <http://www.ebicom.net/~dhyams/cvxpt.htm>. Users are requested to pay a once-only registration fee of US\$40.

5. GLIM

GLIM (Generalized Linear Interactive Modelling, release 4, Aitkin *et al.* 1989) is an old favourite, with which I do most of my analyses. It's compact – the whole system fits on a couple of diskettes. And it is powerful – many analyses can be completed with just a few commands (e.g. Figure 7, in which the data are simulated and analysis completed in one line). But it is not easy to use, so I rarely recommend it to others. There are no pull-down menus, and when GLIM is started, the user is greeted with a simple prompt, as it awaits a command. However, it is flexible and

powerful, and is the only package that I can use to complete some analyses (e.g. Phylogenetic regression, Grafen 1989). The model fitting illustrated below is achieved with the following commands:

```
$yvariate Y
$fit X
$display estimates
$calculate r=%yv-%fv
$plot r %fv
```

The commands (prefixed with a '\$') can be abbreviated, often to one or two characters. The equation $r = \%yv - \%fv$ calculates the residuals by taking the difference between the y-variables (%yv) and the fitted values (%fv). The graph is a plot of the residuals and the fitted values, and is a good way to judge the quality of a fit

GLIM is marketed by the Numerical Algorithms Group in Oxford, UK (NAG Ltd, infodesk@nag.co.uk). When I purchased my copy (several years ago), there was a once-only fee of about £100. Crawley (1993) published a helpful guide to the use of GLIM.

6. ARC

Arc (Cook and Weisberg 1999) is a revision of the program *R-code* (Cook and Weisberg 1994). *R-code* greatly influenced my approach to regression analysis, and *Arc* has become my favourite regression analysis package. It has many innovative tools to allow insightful graphical scrutiny of data.

I quickly found *Arc* intuitive and easy to use. It has pull-down menus, but the user may need to consult the text (Cook and Weisberg 1999) to get the most from the package. Figure 8 shows the analysis of the data used in Figure 6 in *Arc*, but the real strengths of *Arc* become apparent only with more complex sets of data.

ARC is copyrighted software, but may be distributed and used free of charge. It is available from the web at www.stat.umn.edu/arc.

7. VISTA

I've recently discovered ViSta, the Visual Statistics System (Young 2001), and am still learning it, so I'm not well qualified to present it. But I would like to attention to this package, because I think that it is a useful system and it has many innovative features. It encourages graphical examination of data, and has an innovative way to document what has been done, and to suggest what can be done with the data (Figure 9).

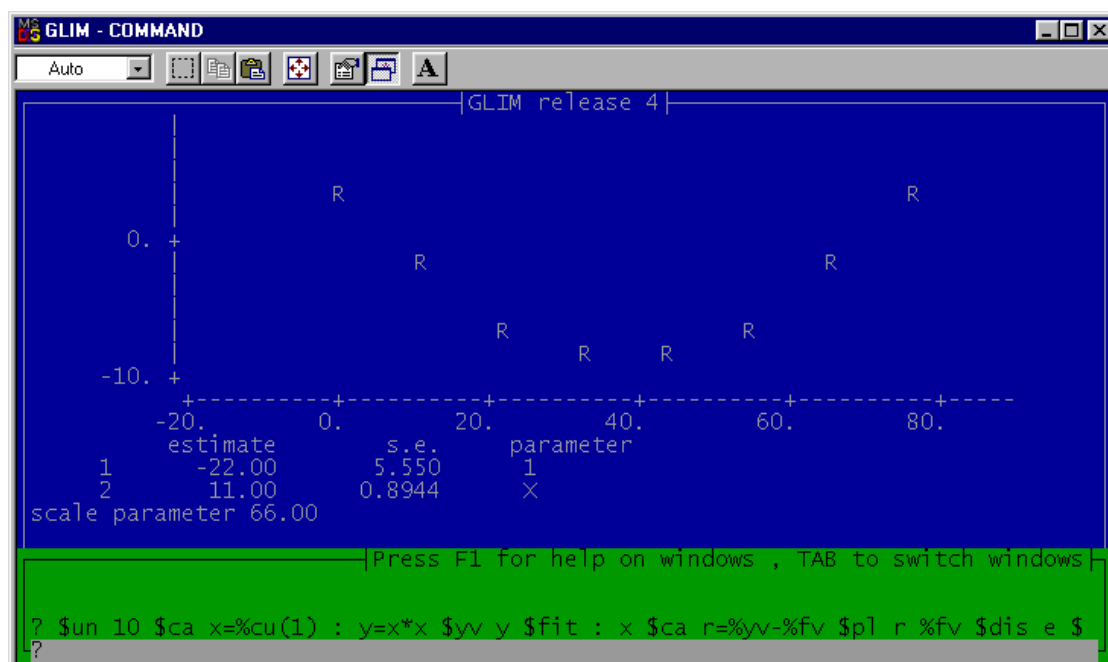


Figure 7. Image of computer screen while using the statistics package GLIM (release 4)

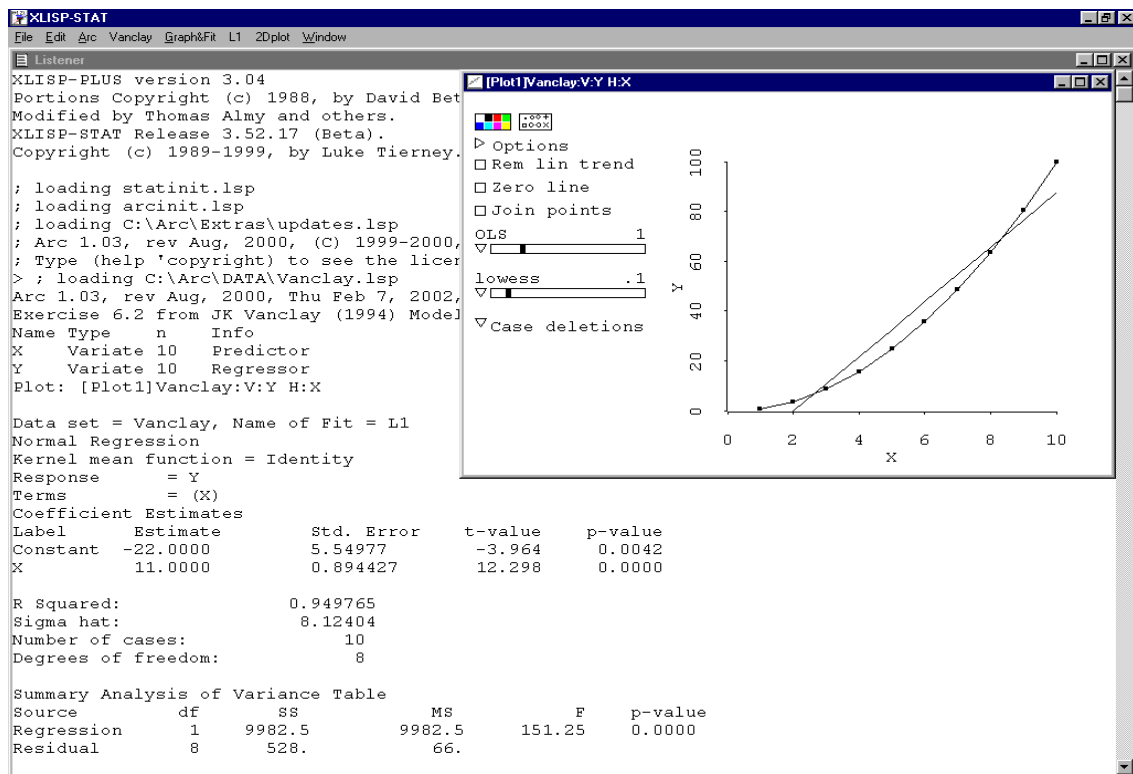


Figure 8. Image of computer screen while using the statistical package ARC

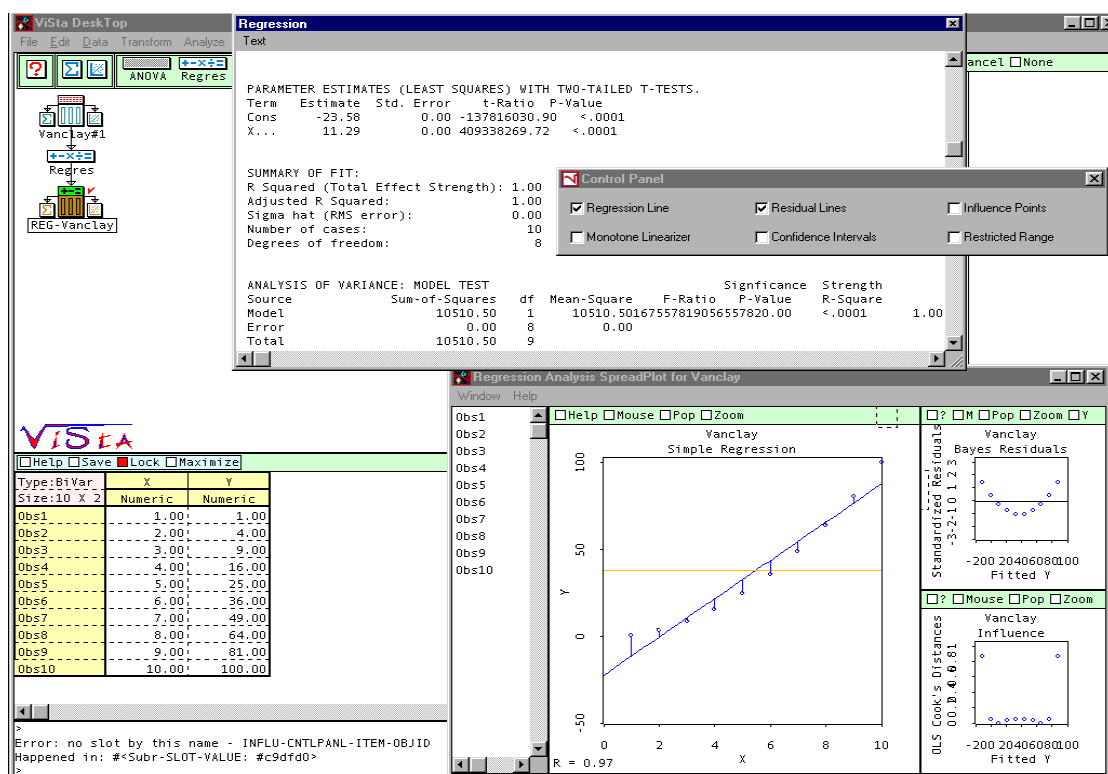


Figure 9. Image of computer screen while using Vista

Note: Note symbols to show what has been done in the analysis (top left), and automatic calculation and display of residuals and influences (bottom right) for every analysis.

ViSta is free, and can be downloaded from the web at www.visualstats.org.

8. CONCLUDING COMMENTS

I am not going to recommend any particular one of these packages – they all have strengths and weaknesses. I like the utility of Excel for data input, the ease of fitting equations CurveExpert, the power of GLIM, the graphics abilities of Arc, and the innovation in ViSta. But Excel has bugs, CurveExpert has limited capabilities, GLIM isn't easy for non-statisticians, Arc doesn't offer sufficient help (unless the text is also obtained), and ViSta takes time to learn. Preferences for particular features and approaches are personal, so provided that candidates have the functionality you need, choose a package that you like. If you don't like the one you have, or find that it cannot handle your data or your analysis, look for another package. There are many good packages out there, and I'm sure that you'll find one that suits your needs and budget.

REFERENCES

- Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989), *Statistical Modelling in GLIM*, Oxford Statistical Science Series 4, Oxford Science Publications, Clarendon Press, Oxford.
- Anscombe, F.J. (1973), 'Graphs in statistical analysis', *American Statistician* 27(1): 17-21.
- Cook, R.D. and Weisberg, S. (1994), *An Introduction to Regression Graphics*, Wiley, New York.
- Cook, R.D. and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics*, Wiley, New York.
- Crawley, M.J. (1993), *GLIM for Ecologists*, Blackwell, Oxford.
- Grafen, A. (1989), 'The phylogenetic regression', *Philosophical Transactions of the Royal Society, London, Series B*, 205, 581-98.
- McCullough, B.D. and Wilson, B. (1999), 'On the accuracy of statistical procedures in Microsoft Excel 97', *Computational Statistics and Data Analysis*, 31(1): 27-37.
- Vanclay, J.K. (1994), *Modelling Forest Growth and Yield: Applications to Mixed Tropical Forests*, CAB International, Wallingford.
- Weisberg, S. (1985), *Applied Linear Regression*, 2nd. edn., Wiley, New York.
- Young, F.W. (2001), How to use ViSta: the Visual Statistics System, The Visual Statistics Project, Psychometric Laboratory, University of North Carolina.